Comprehensive Probability Review

Clase 12 — 14.310x: Data Analysis for Social Scientists

B.Sc. Cristian Lazo Quispe

BREIT (Brescia Institute of Technology) · Aporta — Grupo Breca

SDS Programme (Statistics and Data Science, MITx) · **BREIT** (Brescia Institute of Technology), an Aporta initiative –Grupo Breca– in partnership with MIT IDSS.

clazoq@uni.pe · linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

How does today work?

The format

Today we revisit **probability** (Classes 1–5): counting, conditional probability & Bayes, joint distributions, expectation & variance, and special distributions & the CLT. **15 core problems**, ramping up in difficulty: Easy (3) → Intermediate (6) → Hard (6) — plus **5 bonus problems** at the end in case we have extra time.

What you do

Each slide shows a problem and a blank space: **work it out** there (or on paper) before we move on. No hints, no answer key on these slides.

What I do

For each problem I will read through the **theory used**, the **thought process**, and the **step-by-step solution** out loud, right after we attempt it together.

What we will cover

1. Easy problems (warm-up)
2. Intermediate problems
3. Hard problems (challenge)
4. Bonus problems (if time allows)

Contents

Problem 1 — Easy — Counting (Class 1)

Problem

A district education office is forming a review committee for a school-voucher pilot. From a pool of **7** eligible staff members (all with equal standing, no distinct roles), the office needs to select **3** people to serve on the committee.

Question. How many different 3-person committees can be formed?

Your work

Problem 2 — Easy — Conditional Probability (Class 2)

Problem

A workforce-training program tracks $n = 200$ graduates. The table cross-tabulates whether they completed the full training and whether they are currently employed.

	Employed	Not employed	Total
Completed training	90	30	120
Did not complete	40	40	80
Total	130	70	200

Question. (a) Find $\mathbb{P}(\text{Employed} \mid \text{Completed})$. (b) Find $\mathbb{P}(\text{Employed})$. (c) Are "Employed" and "Completed" independent? Justify using (a) and (b).

Your work

Problem 3 — Easy — Linearity of Expectation (Class 4)

Problem

A vaccination outreach program pays each field worker a fixed weekly stipend of S/. 50 plus a bonus of S/. 8 per household visited. The number of households a worker visits in a week, X, is random with $\mathbb{E}[X] = 18$.

Question. Using linearity of expectation, find the expected weekly payment $\mathbb{E}[50 + 8X]$.

Your work

Contents

Problem 4 — Intermediate — Bayes' Theorem (Class 2)

Problem

A regional health system receives newborns from three types of clinics: urban (50% of all births), peri-urban (30%), and rural (20%). The probability that a newborn requires referral to specialized care is 2% at urban clinics, 5% at peri-urban clinics, and 9% at rural clinics.

Question. A newborn is referred to specialized care. What is the probability that this newborn came from a **rural** clinic?

Your work

Problem 5 — Intermediate — Joint Distributions (Class 3)

Problem

Let X = number of adults in a household (1 or 2) and Y = number of children (0, 1, or 2). Their joint pmf is:

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 1$	0.10	0.15	0.05
$X = 2$	0.10	0.25	0.35

Question. (a) Find the marginal pmf of Y. (b) Are X and Y independent? Justify with one specific probability comparison.

Your work

Problem 6 — Intermediate — Covariance & Variance (Class 4)

Problem

For a portfolio of microloans, let $X =$ monthly revenue growth (%) and $Y =$ monthly expense growth (%) of a borrower's business. Suppose $\mathrm{Var}(X) = 4$, $\mathrm{Var}(Y) = 9$, and $\mathrm{Cov}(X, Y) = -2$ (revenue and expense growth tend to move in opposite directions due to cost-cutting). Define net margin growth as $X - Y$.

Question. Find $\mathrm{Var}(X - Y)$ and its standard deviation.

Your work

Problem 7 — Intermediate — Poisson Distribution (Class 5)

Problem

A rural tele-health call center receives requests for consultations at a rate of $\lambda = 6$ per hour, modeled as a Poisson process.

Question. (a) Find the probability of receiving exactly 4 requests in a given hour. (b) Find the probability of receiving **at least 1** request in a given hour.

Your work

Problem 8 — Intermediate — Counting with a Constraint (Class 1)

Problem

A survey team of **6** distinct volunteers must line up for a group photo. Two of them are the project's co-founders, and they insist on standing **next to each other**.

Question. How many valid arrangements of the 6 volunteers satisfy this condition?

Your work

Problem 9 — Intermediate — CLT & Sample Mean (Class 5)

Problem

Monthly food expenditure per household in a district has population mean S/. 450 and standard deviation S/. 120. A researcher surveys a random sample of $n = 64$ households.

Question. Using the Central Limit Theorem, approximate the probability that the **sample mean** exceeds S/. 470.

Your work

Contents

Problem 10 — Hard — Joint Densities (Class 3)

Problem

An impact evaluation team models how NGO project budgets split between two categories: X = fraction of budget spent on training, Y = fraction spent on materials (the remainder goes to overhead). Since these shares cannot exceed the total budget, (X, Y) is supported on the triangular region $\{x \geq 0,\ y \geq 0,\ x + y \leq 1\}$, with joint density

$$f(x, y) = k\,(x + 2y).$$

Question. (a) Find k. (b) Find the marginal density $f_X(x)$. (c) Find $\mathbb{P}(Y > 2X)$. (d) Find $\mathbb{E}[X]$.

Your work

Problem 11 — Hard — Law of Total Variance (Class 4)

Problem

A ride-hailing platform has two driver pools: experienced drivers (60% of all drivers) and new drivers (40%). Trip completion time T (in minutes) has mean 18 and variance 16 for experienced drivers, and mean 24 and variance 25 for new drivers. A rider is randomly matched to a driver from the platform (without knowing which pool).

Question. Using the laws of total expectation and total variance, find $\mathbb{E}[T]$ and $\mathrm{Var}(T)$ for the completion time of a randomly matched trip.

Your work

Problem 12 — Hard — Sequential Bayesian Updating (Class 2)

Problem

A microfinance officer is assessing default risk for a loan applicant. Prior belief: $\mathbb{P}(\text{High-risk}) = 0.30$, $\mathbb{P}(\text{Low-risk}) = 0.70$. Two screening indicators are checked; conditional on risk type, they are independent of each other. Indicator 1 ("missed a payment-reminder call"): $\mathbb{P}(\text{Yes} \mid \text{High}) = 0.60$, $\mathbb{P}(\text{Yes} \mid \text{Low}) = 0.15$. Indicator 2 ("flagged by credit bureau"): $\mathbb{P}(\text{Yes} \mid \text{High}) = 0.50$, $\mathbb{P}(\text{Yes} \mid \text{Low}) = 0.10$. Both indicators come back "Yes."

Question. Using sequential Bayesian updating (update after Indicator 1, then update again with Indicator 2), find $\mathbb{P}(\text{High-risk} \mid \text{both Yes})$. Then check that your answer matches a direct one-shot Bayes calculation using both signals at once.

Your work

Problem 13 — Hard — MGFs & Sums (Classes 4–5)

Problem

Let $X \sim \text{Exponential}(\lambda)$ model the time (in months) until a microloan defaults, with moment generating function $M_X(t) = \dfrac{\lambda}{\lambda - t}$ for $t < \lambda$.

Question. (a) Use the MGF to derive $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, and $\text{Var}(X)$ in terms of λ. (b) Set $\lambda = 3$. If $T = X_1 + \cdots + X_5$ is the sum of 5 *independent* default clocks (each Exponential(3)), find the MGF of T, name its distribution family, and compute $\mathbb{E}[T]$ and $\text{Var}(T)$.

Your work

Problem 14 — Hard — CLT & Confidence Intervals (Class 5)

Problem

An education nonprofit wants to estimate the proportion p of teachers in a district who have completed a training certification. In a random sample of $n = 225$ teachers, 153 reported having completed it.

Question. (a) Compute the sample proportion $\hat{p}$. (b) Using the CLT, construct an approximate 95% confidence interval for p. (c) Holding $\hat{p}$ fixed, how large would the sample need to be to cut the margin of error in half?

Your work

Problem 15 — Hard — Bayes + Poisson (Capstone, Classes 2 & 5)

Problem

An employment program runs two versions across sites: intensive Version A (65% of sites) and light-touch Version B (35% of sites). At Version A sites, the number of participants finding a job within 3 months per cohort of 10 follows Poisson($\lambda = 7$); at Version B sites, Poisson($\lambda = 4$). A site is picked at random (its version is unknown), and its cohort reports exactly **6** successful placements.

Question. (a) Using Bayes' theorem with the Poisson counts as evidence, find the posterior probability that the site used Version A. (b) Given this observation, what is your posterior-predictive **expected** number of successful placements for a *new* cohort at the same site?

Your work

Contents

Problem 16 — Easy — Counting (Class 1)

Problem

A clinic issues appointment codes using **3 distinct letters** chosen from $\{A, B, C, D, E\}$, arranged in order (e.g., `ABC` and `BAC` are different codes), with no letter repeated.

Question. How many different appointment codes are possible?

Your work

Problem 17 — Intermediate — Independent Events (Class 2)

Problem

In a survey, $\mathbb{P}(\text{uses public transit}) = 0.4$ and $\mathbb{P}(\text{owns a car}) = 0.5$, and the two events are **independent**.

Question. (a) Find $\mathbb{P}(\text{transit and car})$. (b) Find $\mathbb{P}(\text{transit or car})$. (c) Given someone owns a car, what is $\mathbb{P}(\text{transit} \mid \text{car})$?

Your work

Problem 18 — Intermediate — Uniform on a Rectangle (Class 3)

Problem

Let (X, Y) be uniformly distributed on the rectangle $[0, 2] \times [0, 3]$ (constant joint density on the rectangle, 0 outside it).

Question. (a) Find the joint density $f(x, y)$. (b) Find $\mathbb{P}(X < 1 \text{ and } Y < 2)$. (c) Are X and Y independent? (No integration needed for (c) — think about what a *rectangular*, constant-density support implies.)

Your work

Problem 19 — Hard — Best Linear Predictor (Class 4)

Problem

Suppose $\operatorname{Var}(X) = 5$, $\operatorname{Var}(Y) = 12$, and $\operatorname{Cov}(X, Y) = 6$.

Question. (a) Find the slope $b = \operatorname{Cov}(X, Y)/\operatorname{Var}(X)$ of the best linear predictor of Y from X. (b) Find $R^2 = \operatorname{Cov}(X, Y)^2/\big(\operatorname{Var}(X)\operatorname{Var}(Y)\big)$, the share of Y's variance "explained" by that linear predictor.

Your work

Problem 20 — Hard — Sum of Independent Poissons (Class 5)

Problem

Let $X \sim \text{Poisson}(3)$ and $Y \sim \text{Poisson}(5)$ be **independent**.

Question. (a) What is the distribution of $X + Y$, and why? (b) Using that distribution, find $\mathbb{P}(X + Y = 4)$.

Your work

That's a wrap

Recap

15 core problems (plus 5 bonus) across counting, conditional probability & Bayes, joint distributions, expectation & variance, and special distributions & the CLT — the probability core of Classes 1–5.

Let's go through the solutions together.